# **Test Set Accuracy Report for Spam Detection Problem**

# **Using Naïve Bayes and Logistic Regression Algorithms**

## **Naïve Bayes learning algorithm:**

When I keep the vocabulary as is without removing anything, I got accuracy about 0.9079. After removing the basic stop words, the accuracy increases to 0.9205 overall in the test set.

**Command:**

*python .\NaiveBayes\_LogisticRegression.py "NB" "./train/ham" "./train/spam" "./test/ham" "./test/spam" "./stop\_words.txt"*

**Output:**

***##### Running Naive Bayes Learning Algorithm #####***

*With stop words, accuracy = 0.9079497907949791*

*Without stop words, accuracy = 0.9205020920502092*

## **Logistic Regression algorithm:**

When I keep the vocabulary as is without removing anything, I got accuracy around 0.92 for 145 iterations, with learning rate = 0.122 and lambda = 0.007. When removing the basic stop words, the accuracy increases to 0.9288

**Command:**

*python .\NaiveBayes\_LogisticRegression.py "LR" "./train/ham" "./train/spam" "./test/ham" "./test/spam" "./stop\_words.txt" 145 0.122 0.007*

**Output:**

***##### Running Logistic Regression Algorithm #####***

*Number of iterations = 145*

*Learning rate = 0.122*

*Lambda = 0.007*

*With stop words*

*Accuracy after training = 0.9205020920502092*

*###########################*

*Without stop words*

*Accuracy after training = 0.9288702928870293*

If I increase the number of iterations to 500, with learning rate = 0.1, and lambda = 0.001, the initial accuracy = 0.9205. After removing the stop words from the vocabulary, the accuracy increases to 0.9309

**Command:**

*python .\NaiveBayes\_LogisticRegression.py "LR" "./train/ham" "./train/spam" "./test/ham" "./test/spam" "./stop\_words.txt" 500 0.1 0.001*

**Output:**

***##### Running Logistic Regression Algorithm #####***

*Number of iterations = 500*

*Learning rate = 0.1*

*Lambda = 0.001*

*With stop words*

*Accuracy after training = 0.9205020920502092*

*###########################*

*Without stop words*

*Accuracy after training = 0.9309623430962343*

I also generate the iteration, learning rate, and lambda value randomly from Python, also calculate the accuracy after each updated value and record the maximum accuracy. So far the following is the maximum accuracy I can get: 0.935

*##### Running Logistic Regression Algorithm #####*

*Number of iterations = 143*

*Learning rate = 0.12209511250421848*

*Lambda = 0.0072047671136803715*

*With stop words*

*Accuracy after training = 0.9205020920502092*

*###########################*

*Without stop words*

*Accuracy after training = 0.9351464435146444*

The accuracy normally improves after removing the stop words. It may be because these stop words are normally used in both spam and ham emails and has no value in helping classify the spam from regular email. Keeping these stop words may confused the algorithms more as the algorithm think these words help to classify the spam from ham email.